



Soft selective sweeps in complex demographic scenarios

Benjamin A Wilson, Dmitri Petrov and Philipp W Messer

bioRxiv first posted online April 23, 2014

Access the most recent version at doi: <http://dx.doi.org/10.1101/004424>

**Creative
Commons
License**

The copyright holder for this preprint is the author/funder. It is made available under a [CC-BY-NC 4.0 International license](#).

Soft selective sweeps in complex demographic scenarios

Benjamin A. Wilson, Dmitri A. Petrov, Philipp W. Messer

Department of Biology, Stanford University, Stanford, CA 94305

Running title: Soft sweeps and demography

Keywords: adaptation, mutation, coalescent theory

Corresponding author:

Ben Wilson

Department of Biology

Stanford University

371 Serra Mall

Stanford, CA, 94305

phone: +1 650 736 2249

fax: +1 650 723 6132

email: bawilson@stanford.edu

Abstract

Recent studies have shown that adaptation from *de novo* mutation often produces so-called soft selective sweeps, where adaptive mutations of independent mutational origin sweep through the population at the same time. Population genetic theory predicts that soft sweeps should be likely if the product of the population size and the mutation rate towards the adaptive allele is sufficiently large, such that multiple adaptive mutations can establish before one has reached fixation; however, it remains unclear how demographic processes affect the probability of observing soft sweeps. Here we extend the theory of soft selective sweeps to realistic demographic scenarios that allow for changes in population size over time. We first show that population bottlenecks can lead to the removal of all but one adaptive lineage from an initially soft selective sweep. The parameter regime under which such ‘hardening’ of soft selective sweeps is likely is determined by a simple heuristic condition. We further develop a generalized analytical framework, based on an extension of the coalescent process, for calculating the probability of soft sweeps under arbitrary demographic scenarios. Two important limits emerge within this analytical framework: In the limit where population size fluctuations are fast compared to the duration of the sweep, the likelihood of soft sweeps is determined by the harmonic mean of the variance effective population size estimated over the duration of the sweep; in the opposing slow fluctuation limit, the likelihood of soft sweeps is determined by the instantaneous variance effective population size at the onset of the sweep. We show that as a consequence of this finding the probability of observing soft sweeps becomes a function of the strength of selection. Specifically, in species with sharply fluctuating population size, strong selection is more likely to produce soft sweeps than weak selection. Our results highlight the importance of accurate demographic estimates over short evolutionary timescales for understanding the population genetics of adaptation from *de novo* mutation.

Introduction

Adaptation can proceed from standing genetic variation or mutations that are not initially present in the population. When adaptation requires *de novo* mutations, the waiting time until adaptation occurs depends on the product of the mutation rate towards adaptive alleles and the population size. In large populations, or when the mutation rate towards adaptive alleles is high, adaptation can be fast, whereas in small populations the speed of adaptation will often be limited by the availability of adaptive mutations.

Whether adaption is mutation-limited or not has important implications for the population dynamics of adaptive alleles. In a mutation-limited scenario, only a single adaptive mutation typically sweeps through the population and all lineages in a population sample that carry the adaptive allele coalesce into a single ancestor with the adaptive mutation (Figure 1A). This process is referred to as a ‘hard’ selective sweep (HERMISSEON and PENNINGS 2005). Hard selective sweeps leave characteristic signatures in population genomic data, such as a reduction in genetic diversity around the adaptive site (MAYNARD SMITH and HAIGH 1974; KAPLAN *et al.* 1989; KIM and STEPHAN 2002) and the presence of a single, long haplotype (HUDSON *et al.* 1994; SABETI *et al.* 2002; VOIGHT *et al.* 2006). In non-mutation-limited scenarios, by contrast, several adaptive mutations of independent origin can sweep through the population at the same time, producing so-called ‘soft’ selective sweeps (PENNINGS and HERMISSEON 2006a). In a soft sweep, the lineages that carry the adaptive allele collapse into distinct clusters and several haplotypes can be frequent in the population (Figure 1A). As a result, soft sweeps leave more subtle signatures in population genomic data than hard sweeps and are thus more difficult to detect. For example, diversity is not necessarily reduced in the vicinity of the adaptive locus in a soft sweep (PENNINGS and HERMISSEON 2006b).

There is mounting evidence that adaptation is not mutation-limited in many species, even when it requires a specific nucleotide mutation in the genome (MESSER and PETROV 2013). Recent case studies have revealed many examples where, at the same locus, several adaptive mutations of independent mutational origin swept through the population at the same time, producing soft selective sweeps. For instance, soft sweeps have been observed during the evolution of drug resistance in HIV (FISCHER *et al.* 2010; MESSER and NEHER 2012; PENNINGS *et al.* 2014) and malaria (NAIR *et al.* 2007), pesticide and viral resistance in fruit flies (CATANIA *et al.* 2004; AMINETZACH *et al.* 2005; CHUNG *et al.* 2007; KARASOV *et al.* 2010; SCHMIDT *et al.* 2010), warfarin resistance in rats (PELZ *et al.* 2005), and color patterns in beach mice (HOEKSTRA *et al.* 2006; DOMINGUES *et al.* 2012). Even in the global human population, adaptation has produced soft selective sweeps,

as evidenced by the parallel evolution of lactase persistence in Eurasia and Africa through recurrent mutations in the lactase enhancer (BERSAGLIERI *et al.* 2004; TISHKOFF *et al.* 2007; ENATTAH *et al.* 2008; JONES *et al.* 2013) and the mutations in the gene *G6PD* that evolved independently in response to malaria (LOUICHARON *et al.* 2009). Some of these sweeps arose from standing genetic variation while others involved recurrent *de novo* mutation. For the remainder of our study, we will focus on the latter scenario of adaptation arising from *de novo* mutation.

The population genetics of adaptation by soft selective sweeps was first investigated in a series of papers by Hermisson and Pennings (HERMISSON and PENNINGS 2005; PENNINGS and HERMISSON 2006a,b). They found that in a haploid population of constant size the key evolutionary parameter that determines whether adaptation from *de novo* mutations is more likely to produce hard or soft sweeps is the population-scale mutation rate $\Theta = 2N_eU_A$, where N_e is the variance effective population size in a Wright-Fisher model and U_A is the rate at which the adaptive allele arises per individual per generation. When $\Theta \ll 1$, adaptation typically involves only a single adaptive mutation and produces a hard sweep, whereas when Θ becomes on the order of one or larger, soft sweeps predominate (PENNINGS and HERMISSON 2006a).

The strong dependence of the likelihood of soft sweeps on Θ can be understood from an analysis of the involved timescales. An adaptive mutation with selection coefficient s that successfully escapes early stochastic loss requires $\tau_{\text{fix}} \approx \log(N_es)/s$ generations until it eventually fixes in the population (HERMISSON and PENNINGS 2005; DESAI and FISHER 2007). The expected number of independent adaptive mutations that arise during this time is on the order of $NU_A \log(N_es)/s$ – i.e., the product of the population-scale mutation rate towards the adaptive allele and its fixation time. Yet only an approximate fraction $2s$ of these mutations will escape early stochastic loss and successfully establish in the population (HALDANE 1927; KIMURA 1962). Thus, the expected number of independently originated adaptive mutations that successfully establish before the first one has reached fixation is of order $(2s)NU_A \log(N_es)/s = \Theta \log(N_es)$ and, therefore, depends only logarithmically on the selection coefficient of the adaptive allele.

Our current understanding of the likelihood of soft sweeps relies on the assumption of a Wright-Fisher model with fixed population size, where Θ remains constant over time. This assumption is clearly violated in many species, given that population sizes often change dramatically throughout the evolutionary history of a species. In order to assess what type of sweeps to expect in a realistic population, we must understand how the likelihood of soft sweeps is affected by demographic processes.

In many organisms population sizes can fluctuate continuously and over timescales that are not necessarily

long compared to those over which adaptation occurs. For example, many pathogens undergo severe bottlenecks during host-to-host transmission (ARTENSTEIN and MILLER 1966; GERONE *et al.* 1966; WOLFS *et al.* 1992; WANG *et al.* 2010), insects can experience extreme, seasonal boom-bust cycles (WRIGHT *et al.* 1942; IVES 1970; BALTENSWEILER and FISCHLIN 1988; NELSON *et al.* 2013), and even some mammals experience dramatic, cyclical changes in abundance (MYERS 1998; KREBS and MYERS 1974). Extensive work has been devoted to the question of how such fluctuations affect the fixation probabilities of adaptive mutations (EWENS 1967; OTTO and WHITLOCK 1997; POLLAK 2000; PATWA and WAHL 2008; ENGEN *et al.* 2009; PARSONS *et al.* 2010; UECKER and HERMISSON 2011; WAXMAN 2011) but it remains unclear how they affect the likelihood of observing soft sweeps.

In this study we investigate the effects of demographic processes on adaptation from *de novo* mutations. We show that recurrent population bottlenecks can give rise to a phenomenon we term the 'hardening' of soft selective sweeps. Hardening occurs when only one beneficial lineage in an initially soft sweep persists through a population bottleneck. We then develop a generalized analytical framework for calculating the likelihood of soft sweeps under arbitrary demographic scenarios, based on the coalescent with 'killings' process. We find that when population size varies over time, two important symmetries of the constant population size scenario are broken: first, the probability of observing soft sweeps becomes a function of the starting time of the sweep and, second, it becomes a function of the strength of selection. In particular, we show that strong selection is often more likely to produce soft sweeps than weak selection when population size fluctuations are common.

Results

We study a single locus with two alleles, a and A , in a haploid Wright-Fisher population (random mating, discrete generations) (EWENS 2004). The population is initially monomorphic for the wildtype allele a . The derived allele A has a selective advantage s over the wildtype and arises at a rate U_A per individual, per generation. We ignore back mutations and consider the population dynamics of the two alleles at this locus in isolation, *i.e.*, there is no interaction with other alleles elsewhere in the genome.

In a classical hard sweep scenario, a single adaptive allele arises, successfully escapes early stochastic loss, and ultimately sweeps to fixation in the population. In a soft sweep, several adaptive mutations establish independently in the population and rise in frequency before the adaptive allele has fixed in the population. After fixation of the adaptive allele, the lineages in a population sample do not coalesce into a single ancestor

with the adaptive allele but fall into two or more clusters, reflecting the independent mutational origins of the different adaptive lineages (Figure 1A). Note that the distinction between a hard and a soft sweep is based on the genealogy of adaptive alleles in a population sample. It is therefore possible that the same adaptive event yields a soft sweep in one sample but remains hard in another, depending on which individuals are sampled.

Soft sweeps in populations of constant size

The likelihood of soft sweeps during adaptation from *de novo* mutation has been calculated by PENNINGS and HERMISSON (2006a) for a Wright-Fisher model of constant population size N . Using coalescent theory, they showed that in a population sample of size n , drawn right after fixation of the adaptive allele, the probability of observing at least two independently originated adaptive lineages is given by

$$P_{\text{soft},n}(\Theta) \approx 1 - \prod_{k=1}^{n-1} \frac{k}{k + \Theta}, \quad (1)$$

where $\Theta = 2NU_A$ is the population-scale mutation rate – twice the number of adaptive lineages that enter the population per generation. Thus, the probability of a soft sweep is primarily determined by Θ and is nearly independent of the strength of selection.

The transition between the regimes where hard and where soft sweeps predominate occurs when Θ becomes on the order of one in the constant population size scenario. When $\Theta \ll 1$, adaptive mutations are not readily available in the population and adaptation is impeded by the waiting time until the first successful adaptive mutation arises. This regime is referred to as the mutation-limited regime. Adaptation from *de novo* mutation typically produces hard sweeps in this case. When $\Theta \geq 1$, by contrast, adaptive mutations arise at least once per generation on average. In this non-mutation-limited regime, soft sweep predominate.

Soft sweeps under recurrent bottlenecks: heuristic predictions

The standard Wright-Fisher model assumes a population of constant size N . To study the effects of population size changes on the probability of soft sweeps, we relax this condition and model a population that alternates between two sizes. Every ΔT generations the population size is reduced from N_1 to $N_2 \ll N_1$ for a single generation and then returns to its initial size in the following generation (Figure 1B). We define $\Theta = 2N_1U_A$ as the population-scale mutation rate during the large population phases.

We assume instantaneous population size changes and do not explicitly consider a continuous population

decline at the beginning of the bottleneck or growth during the recovery phase. This assumption should be appropriate for sharp, punctuated bottlenecks and allows us to specify the ‘severity’ of a bottleneck in terms of a single parameter, N_2/N_1 . Note that many effects of a population bottleneck depend primarily on the ratio of its duration over its severity. In principle, most of the results we derive below should therefore be readily applicable to more complex bottleneck scenarios by mapping the real bottleneck onto an effective single-generation bottleneck. We also assume that mutation and selection are only operating during the phases when the population is large, whereas the two alleles, a and A , are neutral with respect to each other and no new mutations occur during a bottleneck. This assumption is justified for severe bottlenecks with $N_2 \ll N_1$.

Adaptive mutations arise in the large population at rate $N_1 U_A$, but only a fraction $2s$ of these mutations successfully establishes in the large population, *i.e.*, these mutations stochastically reach a frequency $\approx 1/(N_1 s)$ whereupon they are no longer likely to become lost by random genetic drift (assuming that the amount of drift remains constant over time). Thus, adaptive mutations establish during the large phases at an approximate rate Θs . We assume that successfully establishing mutations reach their establishment frequency fast compared to the timescale ΔT between bottlenecks, in which case establishment can be effectively modeled by a Poisson process. This assumption is reasonable when selection is strong and the establishment frequency low. Note that those adaptive mutations that do reach establishment frequency typically achieve this within very few generations (ERIKSSON *et al.* 2008).

Under the Poisson assumption, the expected waiting time until an adaptive mutation successfully establishes in the large population phase is given by $\tau_{\text{est}} = 1/(\Theta s)$. After establishment, its population frequency is modeled deterministically by logistic growth: $x(t) = 1/[1 + (N_1 s) \exp(-st)]$. Fixation would thus occur within $\tau_{\text{fix}} \approx \log(N_1 s)/s$ generations after establishment, assuming that the population size were to remain constant.

If an adaptive mutation establishes during the large phase but has not yet fixed at the time the next bottleneck occurs, its fate will depend on its frequency at the onset of the bottleneck. In our model, the bottleneck is a single generation of random down-sampling of the population to a size $N_2 \ll N_1$. Any allele present at the onset of the bottleneck will likely survive the bottleneck only when it was previously present at a frequency larger than $1/N_2$, *i.e.*, when at least one copy of the allele is expected to be present during the bottleneck. Less frequent alleles will typically be lost (Figure 1B). To reach frequency $1/N_2$ in the population, an adaptive mutation needs to grow for approximately another $\tau_2 = \log(N_1 s/N_2)/s$ generations after establishment. We can therefore define the ‘bottleneck establishment time’ as the sum of the initial establishment time, τ_{est} , and the waiting time until the mutation has subsequently reached a high-enough

frequency to likely survive a bottleneck, τ_2 :

$$\tau'_{\text{est}} = \frac{1}{\Theta s} + \frac{\log(N_1 s / N_2)}{s}. \quad (2)$$

We will show below that the comparison between bottleneck establishment time, τ'_{est} , and bottleneck recurrence time, ΔT , distinguishes the qualitatively different regimes in our model.

Mutation-limited adaptation: It is clear that bottlenecks can only decrease the probability of a soft sweep in our model relative to the probability in the constant population size scenario, as they systematically remove variation from the population by increasing the variance in allele frequencies between generations. Consequently, when $\Theta \ll 1$ sweeps will be hard because adaptation is already mutation-limited during the large phases. Note that mutation-limitation does not necessarily imply that adaptation is unlikely in general, it may just take longer until an adaptive mutation successfully establishes in the population. When the recurrence time, ΔT , is much larger than the establishment time, τ_{est} , adaptation is still expected to occur between two bottlenecks.

Non-mutation-limited adaptation: If $\Theta \geq 1$, adaptation is not mutation-limited during the large population phases. In the absence of bottlenecks (or when bottlenecks are very weak), adaptation from *de novo* mutation will often produce soft selective sweeps. A strong population bottleneck, however, can potentially remove all but one adaptive lineage and result in a scenario where only this one lineage ultimately fixes. In this case, we say that the bottleneck has ‘hardened’ the initially soft selective sweep.

We can identify the conditions that make hardening likely from a simple comparison of timescales: hardening should occur whenever $\Theta \geq 1$ and at the same time

$$\Delta T < \tau'_{\text{est}}, \quad (3)$$

such that a second *de novo* mutation typically does not have enough time to reach a safe frequency that assures its survival before the next bottleneck sets in (Figure 1B).

The argument that the second adaptive mutation needs to grow for τ_2 generations after its establishment to reach a safe frequency $1/N_2$ only makes sense when the mutation is actually at a lower frequency than $1/N_2$ at establishment, which requires that bottlenecks are sufficiently severe ($N_2/N_1 < s$). For weaker bottlenecks, most established mutations should typically survive the bottleneck and hardening will generally be unlikely. Note that the condition $N_2/N_1 > s$ alone does not imply that soft sweeps should predominate – this still

depends on the value of Θ . In the other limit, where bottleneck severity increases until $N_2 \rightarrow 1$, all sweeps become hardened. This imposes the requirement that $\tau_2 \ll \tau_{\text{fix}}$ or correspondingly that $N_2 \gg 1$ for our bottleneck establishment time to be valid.

The heuristic argument invokes a number of strong simplifications, including that allele frequency trajectories are deterministic once the adaptive allele has reached its establishment frequency, that alleles at frequencies below $1/N_2$ have no chance of surviving a bottleneck, and that establishment occurs instantaneously during a large population phase. In reality, however, an adaptive mutation spends time in the population before establishment. And if this time becomes on the order of ΔT , then adaptive mutations encounter bottlenecks during the process of establishment. In this case, establishment frequency will be higher than $1/(N_1 s)$ and establishment time will be longer than $1/(\Theta s)$ due to the increased drift during bottlenecks. We will address these issues more thoroughly below when we analyze general demographic scenarios.

Our condition relating the bottleneck recurrence time and the bottleneck establishment time (3) makes the interesting prediction that for fixed values of Θ , ΔT , and N_2/N_1 , there should be a threshold selection strength for hardening. Sweeps involving weaker selection than this threshold are likely to be hardened, whereas stronger sweeps are not. Thus, both hard and soft sweeps can occur in the same demographic scenario, depending on the strength of selection. This is in stark contrast to the constant population size scenario, where primarily the value of Θ determines whether adaptation produces hard or soft sweeps while the strength of selection enters only logarithmically.

Soft sweeps under recurrent bottlenecks: forward simulations

We performed extensive forward simulations of adaptation from *de novo* mutation under recurrent population bottlenecks to measure the likelihood of soft sweeps in our model and to assess the accuracy of condition (3) under a broad range of parameter values. In our simulations we modeled the dynamics of adaptive lineages at a single locus in a modified Wright-Fisher model with selection (Methods). To estimate the empirical probability of observing a soft sweep in a given simulation run, we calculated the probability that two randomly sampled adaptive lineages are not identical by descent at the time of fixation of the adaptive allele, *i.e.*, arose from independent mutational origins.

Figure 2 shows phase diagrams of the empirical probabilities of soft sweeps in our simulations over a wide range of parameter values. We investigated three Θ -regimes that differ in the relative proportions at which hard and soft sweeps arise during the large phases before they experience a bottleneck: (i) mostly hard sweeps

arise during the large phase ($\Theta = 0.2$), (ii) mostly soft sweeps arise during the large phase ($\Theta = 2$), and (iii) practically only soft sweeps arise during the large phase ($\Theta = 20$). For each value of Θ , we investigated three different bottleneck severities: $N_1/N_2 = 10^2$, $N_2/N_1 = 10^3$, and $N_1/N_2 = 10^4$.

Our simulations confirm that hardening is common in populations that experience sharp, recurrent bottlenecks. The evolutionary parameters under which hardening is likely are qualitatively distinguished by the heuristic condition (3). Hardening becomes more likely with increasing severity of the population bottlenecks. For a fixed value of Θ and a fixed severity of the bottlenecks, hardening also becomes more likely the weaker the strength of positive selection and the shorter the recurrence time between bottlenecks, as predicted. For the scenarios with $\Theta = 0.2$, most sweeps are already hard when they arise. Thus, there are only few soft sweeps that could be subject to hardening, leading to systematically lower values of P_{soft} compared to the scenarios with higher values of Θ . Note that the transition between the regimes where hardening is common and where it is uncommon can be quite abrupt. For example, in the scenario where $\Theta = 2$, $N_1/N_2 = 10^4$, and $\Delta T = 100$ generations, an adaptive allele with $s = 0.056$ almost always (90%) produced a hard sweep in our simulations, whereas an allele with $s = 0.1$ mostly (57%) produced a soft sweep.

Probability of soft sweeps in complex demographic scenarios

In this section we describe an approach for calculating the probability of observing soft sweeps from recurrent *de novo* mutation that can be applied to complex demographic scenarios. We assume that the population is initially monomorphic for the wildtype allele, a , and that the adaptive allele, A , has selection coefficient s and arises through mutation of the wildtype allele at rate U_A per individual, per generation. Let $P_{\text{soft},n}(t_0, s)$ denote the probability that a sweep arising at time t_0 is soft in a sample of n adaptive alleles. Generally $P_{\text{soft},n}(t_0, s)$ will also be a function of the trajectory, $x(t \geq t_0)$, of the adaptive allele, the specific demographic scenario, $N(t \geq t_0)$, and the sampling time, t_n .

We can calculate $P_{\text{soft},n}(t_0, s)$ given $x(t)$, $N(t)$, and t_n using a straightforward extension of the approach employed by PENNING and HERMISSON (2006a) in deriving $P_{\text{soft},n}(\Theta)$ for a population of constant size, which resulted in Equation (1). In particular, we can model the genealogy of adaptive alleles in a population sample by a coalescent process with ‘killings’ (DURRETT 2008). In this process, two different types of events can occur in the genealogy of adaptive alleles when going backwards in time from the point of sampling: two lineages can coalesce, or a lineage can mutate from the wildtype allele to the adaptive allele (Figure 3). In the latter case, the lineage in which the mutation occurred is stopped (referred to as killing). Thus, each pairwise

coalescence event and each mutation event reduce the number of ancestral lineages in the genealogy by one. The process stops when the last ancestral lineage is stopped by a mutation (which cannot occur further back in the past than time t_0 , the time when the adaptive allele first arose in the population).

Hard and soft sweeps have straightforward interpretations in this framework: In a hard sweep, all lineages in the sample carry the adaptive allele from the same mutational origin and therefore coalesce into a single ancestral lineage before the process finally stops. In a soft sweep, on the other hand, at least one additional mutation occurs before the process stops (Figure 3).

We will depart from the Wright-Fisher framework here and instead model this coalescent as a continuous-time Markov process. The instantaneous rates of coalescence (λ_{coal}) and mutation (λ_{mut}) at time t , assuming that k ancestral lineages are present in the genealogy at this time, are then given by

$$\lambda_{\text{coal}}(t, k) = \frac{k(k-1)}{2N_e(t)x(t)} \quad \text{and} \quad \lambda_{\text{mut}}(t, k) = \frac{kU_A[1-x(t)]}{x(t)}, \quad (4)$$

where $N_e(t)$ is the single-generation variance effective population size in generation t . Note that these are the same rates as used by PENNINGS and HERMISSON (2006a), with the only difference being that in our case the population size $N_e(t)$ can vary over time.

Let us for now assume we were to actually know the times t_1, \dots, t_{n-1} at which coalescence or mutation events happen in the genealogy, where t_k for $k = 1, \dots, n-1$ specifies the time at which the coalescence or mutation event happens that reduces the number of ancestral lineages from $k+1$ to k , and t_n specifies the time of sampling (Figure 3). Note that we do not make any assumptions about when the sample is taken, we only require that there are n copies of the adaptive allele present in the sample. Given a pair of successive time points, t_k and t_{k+1} , we can calculate the probability $P_{\text{coal}}(t_k)$ that this event is a coalescence event, rather than a mutation event, using the theory of competing Poisson processes:

$$\begin{aligned} P_{\text{coal}}(t_k) &= \frac{\int_{t_k}^{t_{k+1}} \lambda_{\text{coal}}(t, k+1) dt}{\int_{t_k}^{t_{k+1}} [\lambda_{\text{coal}}(t, k+1) + \lambda_{\text{mut}}(t, k+1)] dt} \\ &= \frac{k}{k + \Theta_k}. \end{aligned} \quad (5)$$

The last equation holds if we define an effective Θ_k as

$$\Theta_k = 2U_A A_k [(1-x)/x] H_k(N_e x), \quad (6)$$

where $H_k(y) = (t_{k+1} - t_k) / \int_{t_k}^{t_{k+1}} y(t)^{-1} dt$ denotes the harmonic mean and $A_k(y) = \int_{t_k}^{t_{k+1}} y(t) dt / (t_{k+1} - t_k)$ the arithmetic mean, estimated over the interval $[t_k, t_{k+1}]$. This effective Θ_k recovers the original result $\Theta_k = 2N_e U_A$ from PENNING and HERMISSE (2006a) for the special case of constant population size, where $H_k(N_e x) = N_e A_k^{-1}(1/x)$ and mutation and coalescence should only be likely during the early phase of a sweep, when $A_k[(1-x)/x] \approx A_k(1/x)$.

The effective Θ_k from Equation (6) describes the product of two specific means estimated during the time interval between events k and $k+1$: (i) the arithmetic mean of twice the rate at which mutations towards the adaptive allele occur per lineage and (ii) the harmonic mean of $N_e x$, the effective number of individuals that carry the adaptive allele at time t . The first mean is independent of demography and will be largest during the early phase of a sweep when $x(t)$ is small. The second mean depends on the product of both the trajectory, $x(t)$, and the demography, $N_e(t)$. Importantly, as a harmonic mean, it is dominated by the smallest values of $N_e x$ during the estimation interval. Thus, even if the estimation interval lies in a later stage of the sweep, when $x(t)$ is larger than it was early in the sweep, the harmonic mean could nevertheless be small if $N_e(t)$ is small at some point during this interval. In general, when population size varies over time, it is not always true that most coalescence occurs during the early phase of a sweep, and we will therefore not adopt this assumption here. For instance, if a strong bottleneck is encountered late during the sweep, most coalescence can occur within this bottleneck.

Given an arbitrary demographic scenario, $N_e(t)$, and trajectory $x(t)$ of the adaptive allele, Equation (6) allows us to calculate each effective Θ_k if we know the time points t_k and t_{k+1} . Given the sequence $\{\Theta_k\}$ for all $k = 1, \dots, n-1$, we can then calculate the probability that the sweep in our sample is hard, as this is only the case if all individual events in the genealogy happen to be coalescence events. The probability that this happens is the product of all $P_{\text{coal}}(t_k)$. Hence, the probability that the sweep is soft in our sample is

$$\begin{aligned} P_{\text{soft},n}(\{\Theta_k\}) &= 1 - P_{\text{hard},n}(\{\Theta_k\}) \\ &= 1 - \prod_{k=1}^{n-1} \frac{k}{k + \Theta_k}. \end{aligned} \quad (7)$$

Calculating Θ_k for a given demographic scenario

The above calculation of $P_{\text{soft},n}$ based on Equations (6) and (7) presupposed that we actually know the trajectory of the adaptive allele and the times t_k at which coalescence or mutation events occur in the genealogy.

This assumption is unrealistic in practice. A full treatment of the problem in the absence of such information then requires integrating over all possible trajectories and all individual times at which coalescence or mutation events can occur, where we weigh each particular path $x(t)$ and sequence of event times t_1, \dots, t_n by their expected probabilities.

Instead of performing such a complicated ensemble average, we use a deterministic approximation for the trajectory $x(t)$ and then model the times t_k as stochastic random variables that we approximate by their expectation values. Specifically, we model the frequency trajectory of an adaptive allele in the population by

$$x^*(t > t_0) = \frac{e^{s(t-t_0)}}{N(t_0)P_{\text{fix}}(t_0, s)/2 - 1 + e^{s(t-t_0)}}, \quad (8)$$

where $P_{\text{fix}}(t_0, s)$ is the fixation probability of a new mutation of selection coefficient s that arises in the population at time t_0 in a single copy (UECKER and HERMISSON 2011). Calculating such fixation probabilities when population size varies over time has been the subject of several studies and is well understood (EWENS 1967; OTTO and WHITLOCK 1997; POLLAK 2000; PATWA and WAHL 2008; ENGEN *et al.* 2009; PARSONS *et al.* 2010; UECKER and HERMISSON 2011; WAXMAN 2011). For example, UECKER and HERMISSON (2011) have derived the following general formula for calculating $P_{\text{fix}}(t_0, s)$ under arbitrary demographic scenarios:

$$P_{\text{fix}}(t_0, s) = \frac{2}{1 + N(t_0) \int_{t_0}^{\infty} [e^{-s(t-t_0)}/N_e(t)] dt}. \quad (9)$$

Here $N_e(t)$ again specifies the single-generation variance effective population size in generation t . This approximation works well as long as the number of beneficial mutations that enter the population during the sweep is not extremely high ($\Theta \gg 1$), in which case one would need to explicitly include the contribution from mutation in the formulation of the birth-death process.

Assuming that the adaptive allele follows the deterministic trajectory, $x^*(t)$, from Equation (8), we can calculate the expected rates of coalescence, $\lambda_{\text{coal}}^*(t, k)$, and mutation, $\lambda_{\text{mut}}^*(t, k)$, in the genealogy of adaptive alleles in a population sample. Let us assume the sample of size n is taken at t_n . The expectation value $E(t_k)$ of the times t_k ($k = 1, \dots, n-1$) at which the number of lineages goes from $k+1$ to k then obeys the relation

$$n - k = \int_{E(t_k)}^{t_n} [\lambda_{\text{coal}}^*(t, k+1) + \lambda_{\text{mut}}^*(t, k+1)] dt. \quad (10)$$

In other words, $E(t_k)$ is the average waiting time until $n - k$ events have occurred in the genealogy of the

sample when going backwards in time from the point of sampling. Given the expected times $E(t_k)$, one can then calculate the expected values $E(\Theta_k)$ via Equation (6) and estimate $P_{\text{soft},n}(t_0, s)$ via Equation (7).

Application for cycling populations

To illustrate and verify our approach for calculating $P_{\text{soft},n}(t_0, s)$, we examine selective sweeps in a population that undergoes cyclical population size changes. In particular, we model a haploid Wright-Fisher population with a time-dependent population size given by:

$$N(t) = \frac{N_{\min} + N_{\max}}{2} + \frac{N_{\max} - N_{\min}}{2} \sin\left(\frac{2\pi t}{\Delta T}\right). \quad (11)$$

As illustrated in Figure 4A, this specifies a population that cycles between a minimal size, N_{\min} , and a maximal size, N_{\max} , over a period of ΔT generations. We investigate selective sweeps with four different starting times (t_0) at which the successfully sweeping allele first arises within a cycle: $t_0 = 0$, $t_0 = 0.25\Delta T$, $t_0 = 0.5\Delta T$, and $t_0 = 0.75\Delta T$. These four cases describe, in order, a starting time of the sweep midway during a growth phase, at the end of a growth phase, midway of a decline phase, and at the end of a decline phase (Figure 4A). For each starting time we calculate the expected probability $P_{\text{soft},2}(t_0, s)$ of observing a soft sweep in a sample of size two as a function of the selection coefficient (s , of the adaptive allele, assuming that the population is sampled when the adaptive allele has reached population frequency $x = 1/2$). In contrast to sampling at the time of fixation, this criterion does not depend on the actual population size (e.g. in a growing population fixation can take very long). Note that the probability $P_{\text{soft},2}(t_0, s)$ is the probability that two adaptive alleles in a random population sample are not identical by decent.

We derived our analytical predictions for $P_{\text{soft},2}(t_0, s)$ by first calculating $P_{\text{fix}}(t_0, s)$ for the given $N(t)$, t_0 , and s via numerical integration of Equation (9) and then inserting the result into Equation (8) to obtain the trajectory $x^*(t)$, using the scaling $N_e(t) = N(t)/(1 + s)$ for concordance between the generalized birth-death model used by UECKER and HERMISSON (2011) and the Wright-Fisher model. We then estimated $E(t_1)$ via numerical integration of Equation (10) (Methods), assuming that the adaptive allele reaches frequency $x = 1/2$ at:

$$t_2 = t_0 + \frac{\log[N_e(t_0)P_{\text{fix}}(t_0, s)]}{s}. \quad (12)$$

Figure 4B shows the comparison between our analytical predictions for $P_{\text{soft},2}(t_0, s)$ and the observed frequencies of soft sweeps in Wright-Fisher simulations for a scenario with population sizes $N_{\min} = 10^6 =$

$0.01(N_{\max})$, cycle period $\Delta T = 500$, and adaptive mutation rate $U_A = 10^{-8}$, as a function of the strength of positive selection and the starting time of the sweep within a cycle. Simulation results are in good agreement with analytical predictions over the whole range of investigated parameters.

We observe two characteristic limits in our cyclical population size model, specified by the relation between the duration of the sweeps (which inversely depends on the selection strength) and the timescale over which demographic processes occur:

(i) weak selection / fast fluctuation limit: When the duration of a sweep becomes much longer than the period of population size fluctuations, the probability of observing a soft sweep converges to that expected in a population of constant size, given by the harmonic mean of $N_e(t)$ estimated over a population cycle (dash-dotted line in Figure 4B). The starting time of the sweep becomes irrelevant in this case. To show this, we partition the embedded integral $\int 1/(N_e x) dt$ in Equation (6) into consecutive intervals, each extending over one population cycle. Because $x(t)$ changes slow compared with the timescale of a population cycle, we can assume that $x(t)$ is approximately constant over each such interval. The harmonic mean then factorizes into $H_k(N_e x) = H_k(N_e)H_k(x)$, and Equation (6) reduces to

$$\Theta_k = 2U_A H_k(N_e)[1 - H_k(x)] \approx 2U_A H_k(N_e). \quad (13)$$

The last approximation holds as long as k is not too large, in which case the lowest value of $x(t)$ in the interval, and thus also $H_k(x)$, are still small, since the harmonic mean is dominated by the smallest values.

Note that the above argument applies more broadly and is not necessarily limited to scenarios where population size fluctuations are exactly cyclical. In general, a sufficient condition for the factorization in Equation (13) is the existence of a timescale, ξ , that is much shorter than the duration of the sweep, where harmonic averages of $N(t)$ estimated over time intervals of length ξ are already approximately constant for every interval lying within the duration of the sweep. In other words, factorization works for all demographic models that have fast fluctuation modes we can effectively average out but no slow fluctuation modes occurring over timescales comparable to the duration of the sweep.

Examples for demographic models where the weak selection / fast fluctuation limit becomes applicable include those where $N(t)$ is any periodic function with a period much shorter than the duration of the sweep. Another example would be a model in which population sizes are drawn randomly from a distribution with fixed mean, where the number of drawings over the duration of the sweep is large enough such that harmonic

averages already converge to the mean over timescales much shorter than the duration of the sweep.

(ii) strong selection / slow fluctuation limit: When the duration of a sweep becomes much shorter than the timescale over which population size changes, the probability of observing a soft sweep in the cyclical population model converges to that which is expected in a population of constant size $N_e(t_0)$, the effective population size at the starting time of the sweep. In this case the effective Θ_k from Equation (6) reduces to

$$\Theta_k = 2U_A N_e(t_0)[1 - H_k(x)] \approx 2U_A N_e(t_0). \quad (14)$$

We can also recover these weak and strong selection limits for our earlier simulations of the recurrent bottleneck scenario. Figure 4C shows the transition from what is expected in a constant population given by the harmonic mean population size over one bottleneck cycle, $H(N_e)$, to a constant population at the instantaneous population size, $N_e(t_0) = N_1$. The expectations in the limits were calculated using Equation (1) while substituting the appropriate effective population size. Again we see that even for the same demographic scenario, the probability of observing a soft sweep can vary dramatically with selection coefficient. This implies that there is generally no one effective population size that will be relevant for determining the expected selective sweep signature. Notice also that while the transition between the two regimes in our hardening model is monotonic, the transition is not guaranteed to be monotonic in more complex demographic scenarios, as seen for some of the transitions in our cycling population model.

Discussion

In this study we investigated the population parameters that determine the probability of observing soft selective sweeps when adaptation arises from *de novo* mutations. Our understanding of soft sweeps has hitherto been limited to the special case where population size remains constant over time. In this special case, the probability of soft sweeps from recurrent *de novo* mutation depends primarily on the parameter $\Theta = 2N_e U_A$ (twice the population-scale mutation rate towards the adaptive allele) and is largely independent of the strength of selection (PENNING and HERMISSON 2006a). We devised a unified framework for calculating the probability of observing soft sweeps when population size changes over time and found that the strength of selection becomes a key factor for determining the likelihood of observing soft sweeps in many demographic scenarios.

The hardening phenomenon: We first demonstrated that population bottlenecks can give rise to a phenomenon that we term the hardening of soft selective sweeps. Hardening describes a situation where several adaptive mutations of independent origin – initially destined to produce a soft sweep in a constant population – establish in the population, but only one adaptive lineage ultimately survives a subsequent bottleneck, resulting in a hard selective sweep.

Using a simple heuristic approach that models the trajectories of adaptive alleles forward in time, we showed that in populations that experience recurrent, sharp bottlenecks, the likelihood of such hardening depends on the comparison of two characteristic timescales: (i) the recurrence time (ΔT) between bottlenecks and (ii) the bottleneck establishment time (τ'_{est}) which specifies the waiting time until a *de novo* adaptive mutation reaches a high-enough frequency such that it is virtually guaranteed to survive a bottleneck. We derived a simple heuristic approximation, $\tau'_{\text{est}} \approx [\Theta^{-1} + \log(N_1 s / N_2)] / s$, that applies when bottlenecks are severe enough ($N_1 s > N_2$ and $N_2 \gg 1$). If soft sweeps are expected to arise between bottlenecks – *i.e.*, if Θ is on the order of one or larger during those phases – then hardening is common when $\Delta T \ll \tau'_{\text{est}}$, whereas it is unlikely when $\Delta T > \tau'_{\text{est}}$. The bottleneck establishment time increases only logarithmically with the severity of the bottleneck and scales inversely with the selection coefficient of the adaptive mutation. In stark contrast to a population of constant size, the probability of observing soft sweeps can therefore strongly depend on the strength of selection in the recurrent bottleneck scenario.

Generalized analytical framework for complex demographies: The heuristic condition $\Delta T \ll \tau'_{\text{est}}$ provides a rough estimate of whether hardening is expected in a recurrent bottleneck scenario, but it lacks generality for more complex demographic scenarios and does not provide the actual probabilities of observing soft sweeps. We showed that such probabilities can be calculated analytically for a wide range of demographic models by mapping the problem onto a coalescent with killings process (DURRETT 2008). Our approach is very similar to that employed by PENNINGS and HERMISSON (2006a) for the constant size model, with the primary difference being that we allow for coalescence and mutation rates to vary over time as population size changes.

In the coalescent with killings framework (Figure 3), the probability of a soft sweep is determined by the competition between two processes: coalescence of adaptive lineages in the fraction $x(t)$ of the population that carry the adaptive allele, and emergence of new adaptive lineages through *de novo* mutation (referred to as killings when going backwards in time) in the fraction $1 - x(t)$ of the population that do not yet carry the adaptive allele. A sweep is hard in a population sample if all lineages in that sample coalesce before a

second adaptive mutation arises and soft otherwise. In our analytical approach, we assume that the trajectory $x(t)$ can be described by a logistic function. The probability of observing a soft sweep can then be calculated through numerical integration of the expected rates of coalescence and mutation in the genealogy, which are simple functions of $x(t)$ and $N_e(t)$, the variance effective population size in generation t .

Note that by adjusting the end-point of the integration interval to the time at which the adaptive allele reaches a given frequency, our approach can easily be extended to the analysis of partial selective sweeps. Similarly, by extending the time interval beyond the fixation of the adaptive allele, one can study the loss of adaptive lineages due to random genetic drift after the completion of a soft sweep. Moreover, since our model only requires an estimate of the frequency trajectory of the adaptive allele, $x(t)$, it should be easily extendable to other, more complex scenarios, including time-varying selection coefficients (UECKER and HERMISSON 2011), as long as one can still model $x(t)$ in the particular scenario. We leave these possible extensions for future exploration.

Even though the results presented in this paper were derived for haploid populations, it is straightforward to extend them to other levels of ploidy. The key prerequisite is again that we still have an estimate for the frequency trajectory of the adaptive allele, which can be complicated by dominance effects when ploidy increases. Given the trajectory, the population size $N(t)$ simply needs to be multiplied by the ploidy level to adjust for the changed rate of coalescence in the genealogy. For example, in a diploid population with codominance, the population-scale mutation rate needs to be defined as $\Theta = 4N_eU_A$, twice the value for a haploid population of the same size.

Weak and strong selection limits: Our approach reveals interesting analogies to Kingman's coalescent (KINGMAN 1982) with respect to our ability to map the dynamics onto an effective model of constant population size. SJÖDIN *et al.* (2005) showed that genealogies at neutral loci can be described by a linear rescaling of Kingman's coalescent with a corresponding coalescent effective population size, as long as demographic processes and coalescence events operate on very different timescales. Specifically, when population size fluctuations occur much faster compared with the timescale of coalescence, the coalescent effective population size is given by the harmonic mean of the variance effective population size, $N_e(t)$, estimated over the timescale of coalescence. In the opposite limit where population size fluctuations occur much slower than the timescales of neutral coalescence, the variance effective population size is approximately constant over the relevant time interval and directly corresponds to the instantaneous coalescent effective population size.

Analogously, in our analytical framework for determining the likelihood of soft sweeps, we can again map

demography onto an effective model with constant effective population size in the two limits where population size fluctuations are either very fast or very slow. The relevant timescale for comparison here is the duration of the selective sweep, $\tau_{\text{fix}} \approx \log(Ns)/s$, which is inversely proportional to the selection coefficient of the sweep. Hence, the fast fluctuation limit corresponds to a weak selection limit, and the slow fluctuation limit to a strong selection limit. In the strong selection / slow fluctuation limit, the relevant effective population size is the instantaneous effective population size at the start of the sweep; in the weak selection / fast fluctuation limit, it is the harmonic mean of the variance effective population size estimated over the duration of the sweep.

One important consequence of this finding is that, even in the same demographic scenario, the probability of observing soft sweeps can differ substantially for weakly and strongly selected alleles. This is because the harmonic mean that determines the effective population size in the weak selection / fast fluctuation limit will be dominated by the phases where population size is small. For a weakly selected allele in a population that fluctuates much faster than the duration of the sweep, it will be close to the minimum size encountered during the sweep, resulting in a low effective population size and, correspondingly, a low probability of observing a soft sweep. A strongly selected allele, on the other hand, can arise and sweep to fixation between collapses of the population. The effective population size remains large in this case, increasing the probability of observing a soft sweep. Hence, the stronger the selective sweep, the higher the chance that it will be soft in a population that fluctuates in size.

Similar behavior is observed for the fixation probabilities of adaptive alleles in fluctuating populations. In particular, OTTO and WHITLOCK (1997) showed that the fixation process of an adaptive allele depends on the timescale of the fixation itself. Only short-term demographic changes encountered during the fixation event matter for strongly selected alleles, whereas slower changes only affect weakly selected alleles. OTTO and WHITLOCK (1997) therefore concluded that “there is no single effective population size that can be used to determine the probability of fixation for all new beneficial mutations in a population of changing size.”

Hard versus soft selective sweeps in natural populations: How relevant is our finding that the likelihood of observing soft sweeps can strongly depend on the strength of selection for understanding adaptation in realistic populations? We know that both necessary ingredients for this effect to occur – strong temporal fluctuations in population size and difference in the fitness effects of *de novo* adaptive mutations – are common in nature.

Population size fluctuations over several orders of magnitude are observed in various animal species, ranging from parasitic worms to insects and even small mammals (BERRYMAN 2002). Unicellular organisms often

undergo even more dramatic changes in population size. For instance, during Malaria infection only ten to a hundred sporozoites are typically ejected by a feeding mosquito – the numbers of sporozoites that successfully enter the human blood stream are even smaller – yet this population grows to many billions of parasites within an infected individual (ROSENBERG *et al.* 1990). Similarly, in the majority of cases acute HIV infection was found to result from a single virus (KEELE *et al.* 2008). Severe population bottlenecks resulting from serial dilution are also commonly encountered in evolution experiments with bacteria and yeast (WAHL *et al.* 2002). Even our own species has likely experienced population size changes over more than three orders of magnitude within the last 1000 generations (GAZAVE *et al.* 2014).

It is also well established that fitness effects of *de novo* adaptive mutations can vary over many orders of magnitude within the same species. For example, codon bias is typically associated with only weak selective advantages, whereas the fitness advantage during the evolution of drug resistance in pathogens or pesticide resistance in insects can be on the order of 10% or larger.

Taken together, we predict that we should be able to observe strong dependence of the likelihood of hardening on the strength of selection for adaptation in natural populations that experience a demographic phase where adaptation is not mutation-limited. The likelihood of observing soft sweeps will depend on the types of natural population fluctuations that occur and whether they can be characterized by the weak selection / fast fluctuation limit or the strong selection / slow fluctuation limit.

To demonstrate this possibility, consider a cycling population illustrated in Figure 5A that is based on data from the extreme fluctuations observed in multiple species of moths, including the tea tortrix, *Adoxophyes honmai*, and the larch budmoth, *Zeiraphera diniana*. These diploid moth species have been observed to undergo changes in population size spanning many orders of magnitude over short periods of just four to five generations (BALTENSWEILER and FISCHLIN 1988; NELSON *et al.* 2013). Let us further assume that these changes result in a change in the adaptive population-scale mutation rate between $\Theta_{\min} = 10^{-3}$ and $\Theta_{\max} = 1$. In this case, adaptation is not mutation-limited during population maxima and is mutation-limited during population minima. Consequently, hardening of soft selective sweeps is expected to be common.

Figure 5B shows the likelihood of soft sweeps in this scenario according to Equation (7), as a function of the strength of selection and the starting time of the sweep. The probability of observing soft sweeps generally remains low in this scenario, except for cases of extremely strong selection. We can understand this result from the fact that the timescale of population size fluctuations is so fast that all but the most strongly selected alleles still fall within the weak selection limit, described by the harmonic mean.

This result has important consequences for the study of other populations that fluctuate over similarly short timescales, such as the fruit fly *Drosophila melanogaster*. Natural populations of *Drosophila melanogaster* undergo approximately 10–20 generations over a seasonal cycle, often experiencing enormous population sizes during the summer that collapse again each winter (IVES 1970). Our result then suggests that only the most strongly selected alleles, which can arise and sweep over a single season, may actually fall within the strong selection limit. All other sweeps should still be governed by the harmonic mean of the population size averaged over a yearly cycle, which will be dominated by the small winter population sizes.

Let us consider another example, motivated by the proposed recent demographic history of the European human population (COVENTRY *et al.* 2010; NELSON *et al.* 2012; TENNESSEN *et al.* 2012; GAZAVE *et al.* 2014). Specifically, we consider a population that was small throughout most of its history and has recently experienced a dramatic population expansion. We assume demographic parameters similar to those estimated by GAZAVE *et al.* (2014), *i.e.*, an ancestral population size of $N_{\text{anc}} = 10^4$, followed by exponential growth over a period of 113 generations, reaching a current size of approximately $N_{\text{cur}} \approx 520,000$ individuals (Figure 5C). We further assume that exponential growth halts at present and that population size remains constant thereafter. Note that this scenario is qualitatively different from the previously discussed models in that population size changes are non-recurring. As a result, the weak selection / fast fluctuation does not exist in this case. For determining whether a given selective sweep will likely be hard or soft in this model, its starting time becomes of crucial importance.

We assume an adaptive mutation rate of $U_A = 5 \times 10^{-7}$ for this example to illustrate the transition between mutation-limited behavior in the ancestral population, where $\Theta_{\text{anc}} = 4N_{\text{anc}}U_A \approx 0.02$, and non-mutation-limited behavior in the current population, where $\Theta_{\text{cur}} = 4N_{\text{cur}}U_A \approx 1.0$. Note that this adaptive mutation rate is higher than the single nucleotide mutation rate in humans, but it may be appropriate for describing adaptations that have larger mutational target size, such as loss-of-function mutations or changes in the expression level of a gene. Moreover, if we were to assume that the current effective population size of the European human population is in fact $N_{\text{cur}} \approx 2 \times 10^7$ – still over an order of magnitude smaller than its census size – we would already be in the non-mutation-limited regime for $U_A \approx 10^{-8}$, the current estimate of the single nucleotide mutation rate in humans (KONG *et al.* 2012).

Figure 5D shows the probabilities of soft sweeps in this scenario predicted by our approach as a function of the strength of selection and starting time of the sweep. The results confirm our intuition that almost all sweeps that start prior to the expansion are hard in a sample of size two, as expected for adaptation by *de*

novo mutation in a mutation-limited scenario, whereas sweeps starting in the current, non-mutation-limited regime are soft in more than half of the cases, regardless of the strength of selection. Sweeps starting during the expansion phase show an interesting crossover behavior between hard and soft sweeps. The strength of selection becomes important in this case. Specifically, sweeps that start during the expansion have a higher probability of producing soft sweeps when they are driven by weak selection than when they are driven by strong selection. This effect can be understood from the fact that stronger sweeps go to fixation faster than the weaker sweeps. Hence, in a growing population, a weak sweep will experience larger population sizes during its course than a strong sweep starting at the same time, increasing its probability of being soft.

When expanding the intuition from our single-locus model to whole genomes, we must bear in mind that the effective Θ determining the probability of soft sweeps will not be the same for different loci across the genome because mutation rate and target size will vary for adaptive mutations at different loci. For example, adaptive loss-of-function mutations will likely have a much higher value of U_A than adaptive single nucleotide mutations. Therefore, no single value of Θ will be appropriate for describing the entire adaptive dynamics of a population. Adaptation across the genome can simultaneously be mutation-limited and non-mutation-limited in the same population, depending on population size fluctuations, mutation rate, target size, and the strength of selection. Furthermore, we should be very cautious when assuming that estimators for Θ based on genetic diversity will inform us about whether recent adaptation will produce hard or soft sweeps. Estimators based on the levels of neutral diversity in a population, such as Θ_π and Watterson's Θ_W (EWENS 2004), can be strongly biased downward by ancient bottlenecks and recurrent linked selection.

Finally, the overall prevalence of soft sweeps should depend on when adaptation and directional selection is common. If adaptation is limited by mutational input, then most adaptive mutations should arise during the population booms, biasing us toward seeing more soft sweeps. On the other hand, it is also possible – maybe even more probable – that adaptation will be common during periods of population decline, such as when population decline is caused by a strong selective agent like a new pathogen, competitor, predator, or a shortage in the abundance of food. If adaptation is more common during population busts, this should lead us to observe more hard sweeps.

These considerations highlight one of the key limits of the current analysis – we have only considered scenarios where population size and selection coefficients are independent of each other. In the future, we believe that models that consider population size and fitness in a unified framework will be necessary to fully understand signatures that adaptation leaves in populations of variable size.

Methods

Forward simulations of adaptation under recurrent population bottlenecks: We simulated adaptation from *de novo* mutation in a modified Wright-Fisher model with selection. Each simulation run was started from a population that was initially monomorphic for the wildtype allele, a . New adaptive mutations entered the population by a Poisson process with rate $N_1 U_A [1 - x(t)]$, where $1 - x(t)$ is the frequency of the wildtype allele. The population in each generation was produced by multinomial sampling from the previous generation, with sampling probabilities being proportional to the difference in fitness of each lineage and the mean population fitness. Population bottlenecks were simulated through a single-generation downsampling to size N_2 (without selection) every ΔT generations. We did not require that the first beneficial mutation arise in the first generation. Each simulation run started ΔT generations before the first bottleneck. All adaptive lineages were tracked in the population until the adaptive allele had reached fixation. One thousand simulations were run for each parameter combination. Empirical probabilities of observing a soft sweep in a given simulation run were obtained by calculating the expected probability that two randomly drawn adaptive lineages are not identical by descent, based on the population frequencies of all adaptive lineages in the population at the time of sampling. All code was written in Python and C++ and is available upon request.

Numerical Monte Carlo integration: Analytical predictions for $P_{\text{soft},2}(t, s)$ in Figure 4 and Figure 5 were obtained by the following procedure: For the given demographic model, selection coefficient, and starting time of the sweep, we first calculated the fixation probability of the adaptive allele via Equation (9) using Monte Carlo integration routines from the GNU Scientific Library (GALASSI *et al.* 2009). This fixation probability was then used in Equation (8) to obtain the deterministic trajectory $x^*(t)$. Solving $x^*(t_2) = 1/2$ yielded the sampling time t_2 . We then iteratively estimated the lower bound $E(t_1)$ of integral (10) such that the expected number of events occurring between $E(t_1)$ and t_2 converged to 1 ± 10^{-4} . Finally, we integrated the coalescence rate from Equation (4) over the interval $[E(t_1), t_2]$ to determine the probability that the event occurring at $E(t_1)$ was a coalescent event, yielding $P_{\text{coal}} = 1 - P_{\text{soft},2}$. Note that this approach can easily be adjusted for any other sampling time or adaptive allele frequency at sampling.

Forward simulations in cycling populations: We simulated adaptation from *de novo* mutation in a cycling population using the Wright-Fisher model specified above. Each simulated population was initially monomorphic for the wildtype allele. We began our simulations at four different time points (t_0) along the population growth cycle and ran each simulation on the condition that the first beneficial allele that arose in generation

t_0 did not go extinct during the simulation. Simulations were run until the adaptive allele was above 50% frequency. Ten thousand simulations were run for each combination of parameters.

Acknowledgments

We thank Pleuni Pennings, Jamie Blundell, and Hildegard Uecker for useful discussions leading to the formulation of our primary results. We thank Nandita Garud, Joachim Hermisson, Marc Feldman, Daniel Fisher, and members of the Petrov lab for comments and suggestions made prior to and during the formulation of this manuscript. B.A.W. is supported by the NSF Graduate Research Fellowship. This work was supported by the NIH under grants GM089926 and HG002568 to D.A.P.

Literature Cited

- AMINETZACH, Y. T., J. M. MACPHERSON, and D. A. PETROV, 2005 Pesticide resistance via transposition-mediated adaptive gene truncation in *Drosophila*. *Science* **309**: 764–767.
- ARTENSTEIN, M. S., and W. S. MILLER, 1966 Air sampling for respiratory disease agents in army recruits. *Bacteriological reviews* **30**: 571.
- BALTENSWEILER, W., and A. FISCHLIN, 1988 The larch budmoth in the Alps. In *Dynamics of forest insect populations*. Springer, 331–351.
- BERRYMAN, A., 2002 *Population Cycles: The Case for Trophic Interactions*. Oxford University Press.
- BERSAGLIERI, T., P. C. SABETI, N. PATTERSON, T. VANDERPLOEG, S. F. SCHAFFNER, *et al.*, 2004 Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74**: 1111–1120.
- CATANIA, F., M. KAUER, P. DABORN, J. YEN, R. FFRENCH-CONSTANT, *et al.*, 2004 World-wide survey of an Accord insertion and its association with DDT resistance in *Drosophila melanogaster*. *Molecular ecology* **13**: 2491–2504.
- CHUNG, H., M. R. BOGWITZ, C. MCCART, A. ANDRIANOPOULOS, R. H. FFRENCH-CONSTANT, *et al.*, 2007 Cis-regulatory elements in the Accord retrotransposon result in tissue-specific expression of the *Drosophila melanogaster* insecticide resistance gene *Cyp6g1*. *Genetics* **175**: 1071–1077.

- COVENTRY, A., L. M. BULL-OTTERSON, X. LIU, A. G. CLARK, T. J. MAXWELL, *et al.*, 2010 Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature communications* **1**: 131.
- DESAI, M. M., and D. S. FISHER, 2007 Beneficial mutation selection balance and the effect of linkage on positive selection. *Genetics* **176**: 1759–1798.
- DOMINGUES, V. S., Y. P. POH, B. K. PETERSON, P. S. PENNINGS, J. D. JENSEN, *et al.*, 2012 Evidence of adaptation from ancestral variation in young populations of beach mice. *Evolution* **66**: 3209–3223.
- DURRETT, R., 2008 *Probability models for DNA sequence evolution*. Springer.
- ENATTAH, N. S., T. G. JENSEN, M. NIELSEN, R. LEWINSKI, M. KUOKKANEN, *et al.*, 2008 Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *Am. J. Hum. Genet.* **82**: 57–72.
- ENGEN, S., R. LANDE, and B.-E. SÆTHER, 2009 Fixation probability of beneficial mutations in a fluctuating population. *Genetics research* **91**: 73–82.
- ERIKSSON, A., P. FERNSTROM, B. MEHLIG, and S. SAGITOV, 2008 An accurate model for genetic hitchhiking. *Genetics* **178**: 439–451.
- EWENS, W. J., 1967 The probability of survival of a new mutant in a fluctuating environment. *Heredity* **22**: 438–443.
- EWENS, W. J., 2004 *Mathematical Population Genetics*. Springer, New York, 2nd edition.
- FISCHER, W., V. V. GANUSOV, E. E. GIORGI, P. T. HRABER, B. F. KEELE, *et al.*, 2010 Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS ONE* **5**: e12303.
- GALASSI, M., J. DAVIES, J. THEILER, B. GOUGH, G. JUNGMAN, *et al.*, 2009 *GNU scientific library: Reference manual*. Network Theory, Bristol, UK, 3rd edition.
- GAZAVE, E., L. MA, D. CHANG, A. COVENTRY, F. GAO, *et al.*, 2014 Neutral genomic regions refine models of recent rapid human population growth. *Proc. Natl. Acad. Sci. U.S.A.* **111**: 757–762.

- GERONE, P. J., R. B. COUCH, G. V. KEEFER, R. DOUGLAS, E. B. DERRENBACHER, *et al.*, 1966 Assessment of experimental and natural viral aerosols. *Bacteriological reviews* **30**: 576.
- HALDANE, J. B. S., 1927 A mathematical theory of natural and artificial selection, part V: Selection and mutation. *Mathematical Proceedings of the Cambridge Philosophical Society* **23**: 838–844.
- HERMISSON, J., and P. S. PENNINGS, 2005 Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**: 2335–2352.
- HOEKSTRA, H. E., R. J. HIRSCHMANN, R. A. BUNDEY, P. A. INSEL, and J. P. CROSSLAND, 2006 A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science* **313**: 101–104.
- HUDSON, R. R., K. BAILEY, D. SKARECKY, J. KWIATOWSKI, and F. J. AYALA, 1994 Evidence for positive selection in the superoxide dismutase (Sod) region of *Drosophila melanogaster*. *Genetics* **136**: 1329–1340.
- IVES, P. T., 1970 Further genetic studies of the South Amherst population of *Drosophila melanogaster*. *Evolution* **24**: 507–518.
- JONES, B. L., T. O. RAGA, A. LIEBERT, P. ZMARZ, E. BEKELE, *et al.*, 2013 Diversity of lactase persistence alleles in Ethiopia: Signature of a soft selective sweep. *The American Journal of Human Genetics* **93**: 538–544.
- KAPLAN, N. L., R. R. HUDSON, and C. H. LANGLEY, 1989 The "hitchhiking effect" revisited. *Genetics* **123**: 887–899.
- KARASOV, T., P. W. MESSER, and D. A. PETROV, 2010 Evidence that adaptation in *Drosophila* is not limited by mutation at single sites. *PLoS Genet.* **6**: e1000924.
- KEELE, B. F., E. E. GIORGI, J. F. SALAZAR-GONZALEZ, J. M. DECKER, K. T. PHAM, *et al.*, 2008 Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc. Natl. Acad. Sci. U.S.A.* **105**: 7552–7557.
- KIM, Y., and W. STEPHAN, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**: 765–777.
- KIMURA, M., 1962 On the probability of fixation of mutant genes in a population. *Genetics* **47**: 713–719.

- KINGMAN, J., 1982 The coalescent. *Stochastic Processes and their Applications* **13**: 235–248.
- KONG, A., M. L. FRIGGE, G. MASSON, S. BESENBACHER, P. SULEM, *et al.*, 2012 Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**: 471–475.
- KREBS, C. J., and J. H. MYERS, 1974 Population cycles in small mammals. *Advances in ecological research* **8**: 267–399.
- LOUICHAROEN, C., E. PATIN, R. PAUL, I. NUCHPRAYOON, B. WITOONPANICH, *et al.*, 2009 Positively selected G6PD-Mahidol mutation reduces *Plasmodium vivax* density in Southeast Asians. *Science* **326**: 1546–1549.
- MAYNARD SMITH, J., and J. HAIGH, 1974 The hitch-hiking effect of a favourable gene. *Genetical Research* **23**: 23–35.
- MESSER, P. W., and R. A. NEHER, 2012 Estimating the strength of selective sweeps from deep population diversity data. *Genetics* **191**: 593–605.
- MESSER, P. W., and D. A. PETROV, 2013 Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol. Evol.* **28**: 659–669.
- MYERS, J. H., 1998 Synchrony in outbreaks of forest Lepidoptera: a possible example of the Moran effect. *Ecology* **79**: 1111–1117.
- NAIR, S., D. NASH, D. SUDIMACK, A. JAIDEE, M. BARENDSE, *et al.*, 2007 Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites. *Mol. Biol. Evol.* **24**: 562–573.
- NELSON, M. R., D. WEGMANN, M. G. EHM, D. KESSNER, P. S. JEAN, *et al.*, 2012 An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**: 100–104.
- NELSON, W. A., O. N. BJRNSTAD, and T. YAMANAKA, 2013 Recurrent insect outbreaks caused by temperature-driven changes in system stability. *Science* **341**: 796–799.
- OTTO, S. P., and M. C. WHITLOCK, 1997 The probability of fixation in populations of changing size. *Genetics* **146**: 723–733.

- PARSONS, T. L., C. QUINCE, and J. B. PLOTKIN, 2010 Some consequences of demographic stochasticity in population genetics. *Genetics* **185**: 1345–1354.
- PATWA, Z., and L. WAHL, 2008 The fixation probability of beneficial mutations. *Journal of The Royal Society Interface* **5**: 1279–1289.
- PELZ, H. J., S. ROST, M. HUNERBERG, A. FREGIN, A. C. HEIBERG, *et al.*, 2005 The genetic basis of resistance to anticoagulants in rodents. *Genetics* **170**: 1839–1847.
- PENNINGS, P. S., and J. HERMISSON, 2006a Soft sweeps II—molecular population genetics of adaptation from recurrent mutation or migration. *Mol. Biol. Evol.* **23**: 1076–1084.
- PENNINGS, P. S., and J. HERMISSON, 2006b Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS Genet.* **2**: e186.
- PENNINGS, P. S., S. KRYAZHIMSKIY, and J. WAKELEY, 2014 Loss and recovery of genetic diversity in adapting populations of HIV. *PLoS Genet.* **10**: e1004000.
- POLLAK, E., 2000 Fixation probabilities when the population size undergoes cyclic fluctuations. *Theoretical population biology* **57**: 51–58.
- ROSENBERG, R., R. A. WIRTZ, I. SCHNEIDER, and R. BURGE, 1990 An estimation of the number of malaria sporozoites ejected by a feeding mosquito. *Trans. R. Soc. Trop. Med. Hyg.* **84**: 209–212.
- SABETI, P. C., D. E. REICH, J. M. HIGGINS, H. Z. LEVINE, D. J. RICHTER, *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**: 832–837.
- SCHMIDT, J. M., R. T. GOOD, B. APPLETON, J. SHERRARD, G. C. RAYMANT, *et al.*, 2010 Copy number variation and transposable elements feature in recent, ongoing adaptation at the Cyp6g1 locus. *PLoS Genet.* **6**: e1000998.
- SJÖDIN, P., I. KAJ, S. KRONE, M. LASCOUX, and M. NORDBORG, 2005 On the meaning and existence of an effective population size. *Genetics* **169**: 1061–1070.
- TENNESSEN, J. A., A. W. BIGHAM, T. D. OCONNOR, W. FU, E. E. KENNY, *et al.*, 2012 Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**: 64–69.

- TISHKOFF, S. A., F. A. REED, A. RANCIARO, B. F. VOIGHT, C. C. BABBITT, *et al.*, 2007 Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* **39**: 31–40.
- UECKER, H., and J. HERMISSON, 2011 On the fixation process of a beneficial mutation in a variable environment. *Genetics* **188**: 915–930.
- VOIGHT, B. F., S. KUDARAVALLI, X. WEN, and J. K. PRITCHARD, 2006 A map of recent positive selection in the human genome. *PLoS Biol.* **4**: e72.
- WAHL, L. M., P. J. GERRISH, and I. SAIKA-VOIVOD, 2002 Evaluating the impact of population bottlenecks in experimental evolution. *Genetics* **162**: 961–971.
- WANG, G. P., S. A. SHERRILL-MIX, K.-M. CHANG, C. QUINCE, and F. D. BUSHMAN, 2010 Hepatitis C virus transmission bottlenecks analyzed by deep sequencing. *Journal of virology* **84**: 6218–6228.
- WAXMAN, D., 2011 A unified treatment of the probability of fixation when population size and the strength of selection change over time. *Genetics* **188**: 907–913.
- WOLFS, T. F., G. ZWART, M. BAKKER, and J. GOUDSMIT, 1992 HIV-1 genomic RNA diversification following sexual and parenteral virus transmission. *Virology* **189**: 103–110.
- WRIGHT, S., T. DOBZHANSKY, and W. HOVANITZ, 1942 Genetics of natural populations. VII. the allelism of lethals in the third chromosome of *Drosophila pseudoobscura*. *Genetics* **27**: 363.

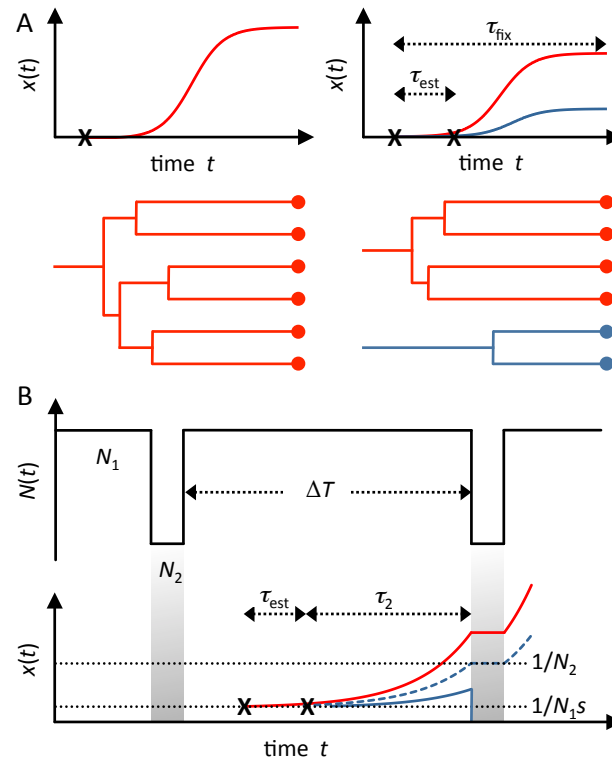


Figure 1: Hard and soft sweeps in populations of constant size and under recurrent population bottlenecks. (A) Allele frequency trajectories and corresponding coalescent genealogies for a hard selective sweep (left) and a soft selective sweep (right). In the soft sweep scenario, a second beneficial mutation establishes τ_{est} generations after the first mutation but before the beneficial allele has fixed. The distinguishing feature between a hard and a soft sweep can be seen in the genealogy of a population sample of individuals with the adaptive allele: in a hard sweep, the sample coalesces into a single ancestor, whereas in a soft sweep the sample coalesces into multiple ancestors with independently arisen adaptive mutations. (B) Illustration of our simplified model used to explore the hardening phenomenon. Population bottlenecks occur every ΔT generations wherein the population size is reduced from N_1 to N_2 for a single generation. The average waiting time between independently establishing beneficial mutations is τ_{est} . From establishment, it takes τ_2 generations for the second mutation to reach frequency $1/N_2$, from where on it is unlikely to be lost during the bottleneck. The hardening phenomenon is illustrated by the loss of the dark blue allele during the bottleneck. The dashed blue line indicates the threshold trajectory required for the mutation to successfully survive the bottleneck.

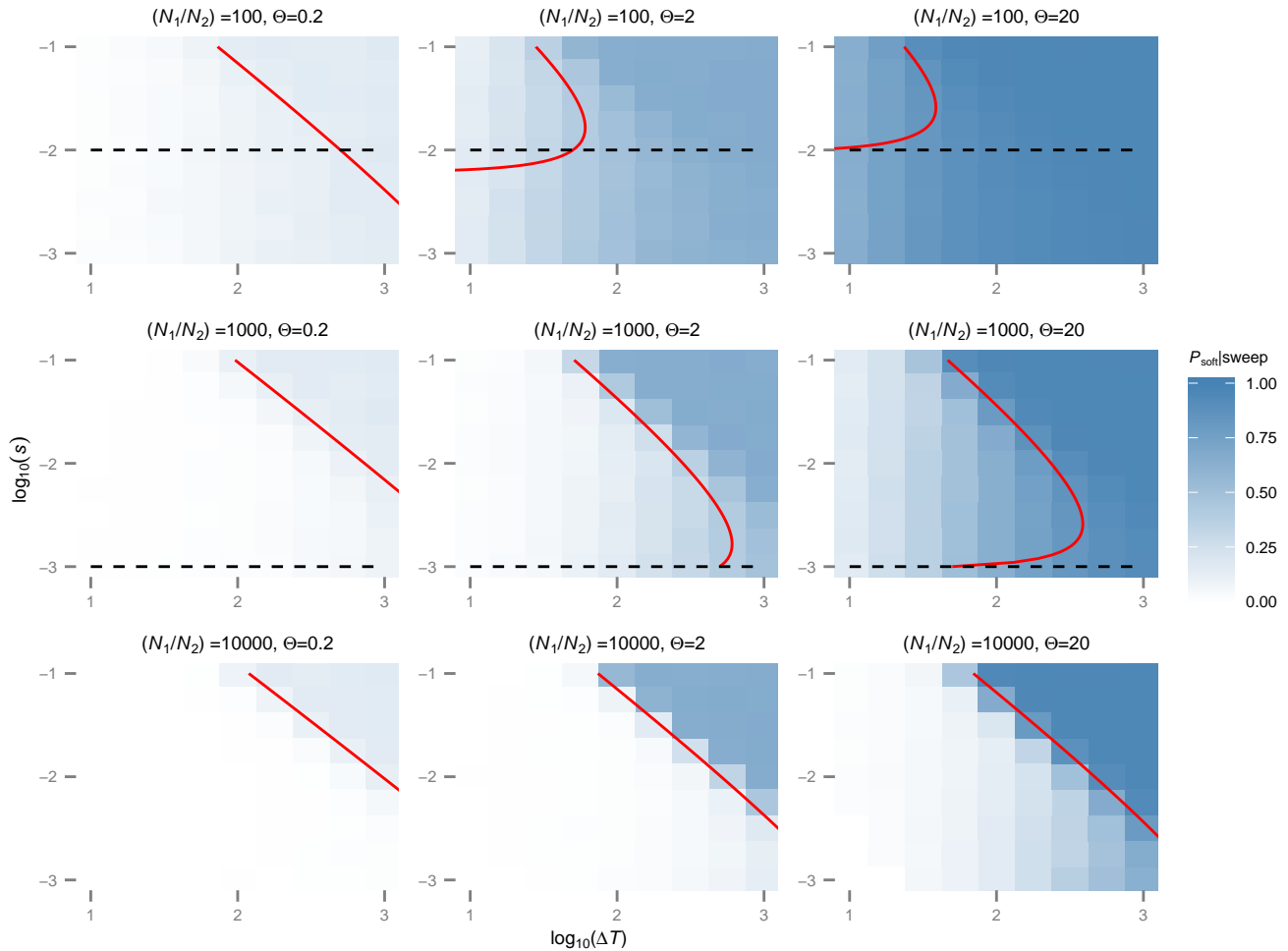


Figure 2: Hardening of soft selective sweeps under recurrent population bottlenecks. The nine panels show different bottleneck severities (weaker to stronger from top to bottom) and different population-scale mutation rates, Θ , during the large population phases. The coloring of the squares specifies the proportion of soft sweeps observed in samples of two individuals at the time of fixation for 1000 simulations runs (Methods) with selection coefficient (s) and bottleneck recurrence time (ΔT) at the center of each square. The red lines indicate the boundary condition $\Delta T = \tau'_{\text{est}}$ between the regime where hardening is predicted to be likely (left of line) and unlikely (right of line) according to our heuristic Equation (3). The dashed black line indicates the boundary condition $N_2/N_1 = s$ on the severity of the bottleneck; below the line, bottlenecks are not severe enough for the hardening condition to be applicable. Note that for the low population-scale mutation rate $\Theta = 0.2$ in the left panels only very few sweeps are soft initially during the large population phase, and hardening therefore is unlikely from the outset.

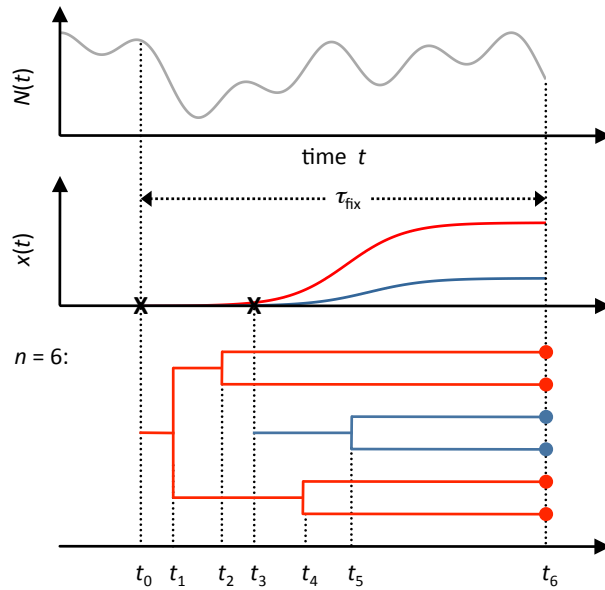


Figure 3: Modeling the genealogy of adaptive alleles by a coalescent process with killings. Population size $N(t)$ can vary arbitrarily over time in our model (top panel). An adaptive allele arises in the population (indicated by \times) in generation t_0 and subsequently sweeps through the population (red frequency trajectory $x(t)$ in middle panel). Before fixation, a second adaptive lineage arises by mutation (indicated by second \times) and also sweeps through the population (blue frequency trajectory in middle panel). The bottom panel illustrates a possible genealogy of a population sample of $n = 6$ adaptive alleles, taken at the time t_6 . When tracing the lineages back in time, a pair of lineages can coalesce (events t_1, t_2, t_4 , and t_5) or a lineage can mutate (events t_0 and t_3), indicating *de novo* mutational origin of the adaptive allele. In the latter case the lineage is killed. The shown example is a soft sweep because a second *de novo* mutation occurs before all lineages have coalesced into a single ancestor.

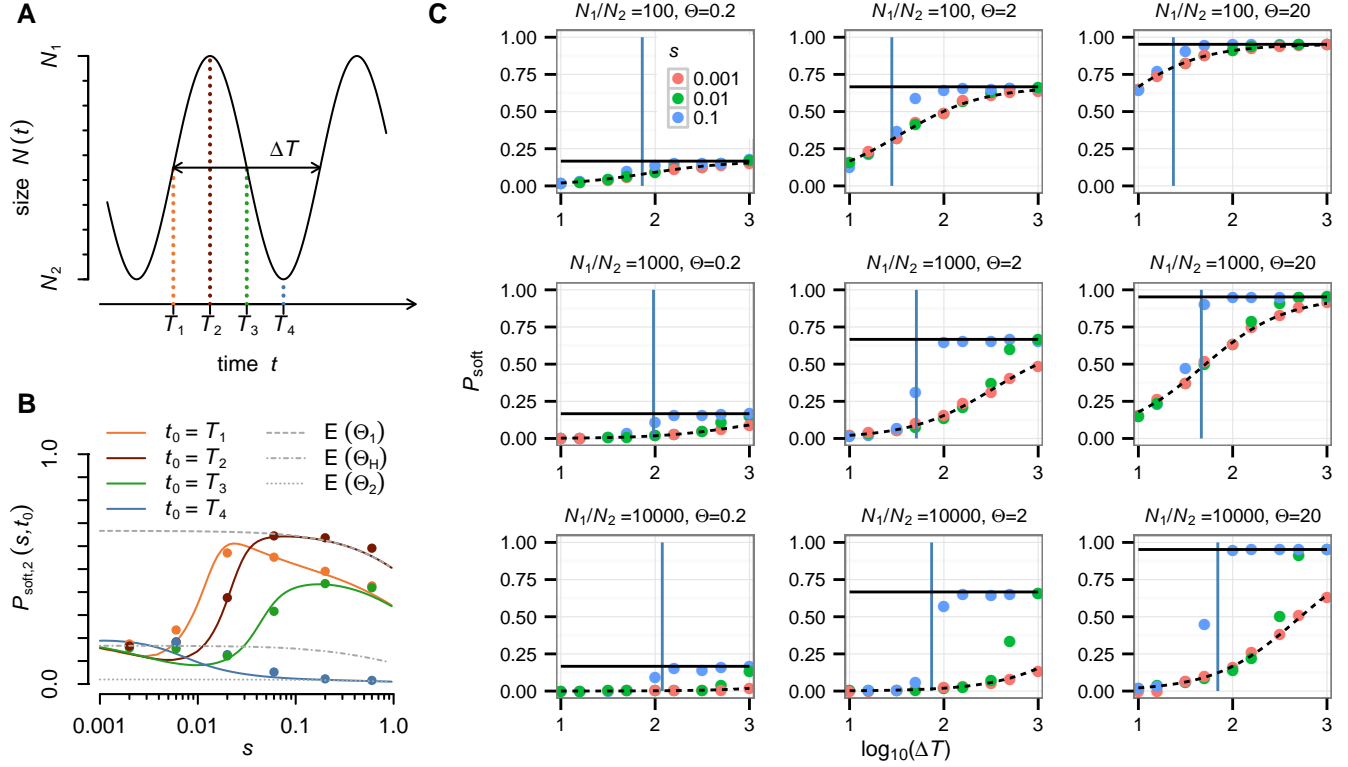


Figure 4: Weak and strong selection limits. (A) In the cyclical population model, $N(t)$ cycles between a maximum size $N_{\max} = 10^8$ and a minimum size $N_{\min} = 10^6$ over a period of $\Delta T = 500$ generations. Adaptive mutations occur at a *de novo* rate of $U_A = 10^{-8}$ per individual, per generation. We condition selective sweeps on four different starting times: T_1 , T_2 , T_3 , and T_4 . (B) Comparison of our analytical predictions for the probabilities $P_{\text{soft},2}(t)$ of observing a soft sweep in a sample of two adaptive alleles, drawn randomly at the time when the adaptive allele has reached a population frequency of 50% (colored lines), with empirical probabilities observed in Wright-Fisher simulations (colored circles, see Methods). Convergence to the harmonic mean expectation, $E(\Theta_H)$, is seen for weak selection, while convergence to the instantaneous population size expectation, $E(\Theta_1)$ and $E(\Theta_2)$, is seen for strong selection. The convergence of the orange and light blue lines is also expected in the strong selection limit as they share the same instantaneous population size at t_0 . (C) The weak selection / fast fluctuation and strong selection / slow fluctuation limits are also observed in our recurrent bottleneck model from Figure 2. The observed probabilities of soft sweeps in the recurrent bottleneck simulations transition from the harmonic mean expectations (dashed black lines) to the instantaneous population size expectations (solid black lines). The solid blue vertical line indicates the position of our heuristic boundary $\Delta T = \tau'_{\text{est}}$ for the selection coefficient $s = 0.1$.

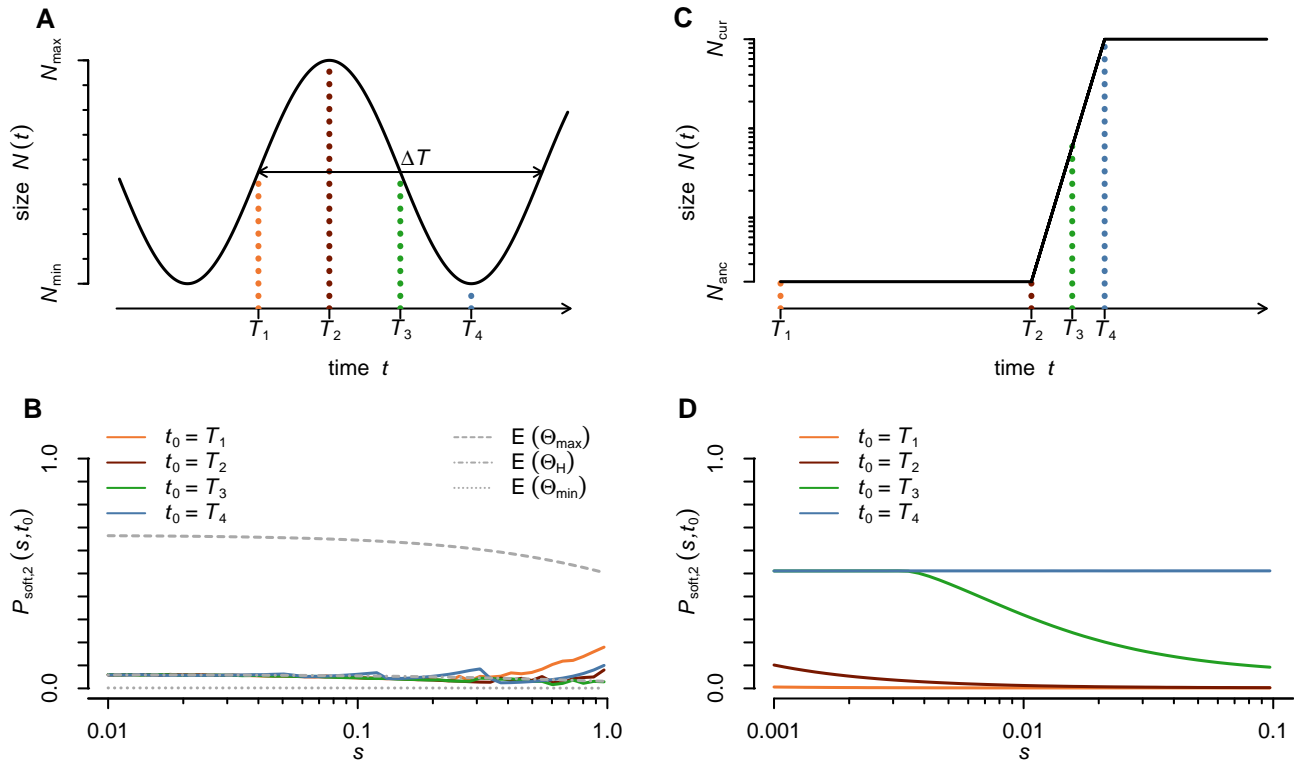


Figure 5: Probability of observing soft sweeps in two demographic scenarios. (A) Fluctuating population inspired by data from the extreme fluctuations observed in multiple species of moths (BALTENSWEILER and FISCHLIN 1988; NELSON *et al.* 2013). We assume the adaptive population-scale mutation rate varies between $\Theta_{\min} = 10^{-3}$ and $\Theta_{\max} = 1$ over a period of $\Delta T = 5$ generations. (B) Our predictions for the probability of observing a soft sweep in a sample of two adaptive alleles drawn randomly at the time-point when the adaptive allele has reached a population frequency of 50%, conditional on four different starting times of the sweep (T_1 to T_4). The noise in the predictions stems from the numerical Monte Carlo integrations in our approach. The probability of observing a soft sweep is close to the harmonic mean expectation, $E(\Theta_H)$, for virtually all starting times and selection strengths, except when selection is extremely strong. (C) Demographic model proposed for the European human population (GAZAVE *et al.* 2014). The ancestral population size is $N_{\text{anc}} = 10^4$. Starting at 113 generations in the past, the population expands exponentially at a constant rate of $r = 0.0554$, until it reaches its current size of $N_{\text{cur}} \approx 520,000$. Population size is assumed to remain constant thereafter. Note that the y-axis is plotted logarithmically. We set the beneficial mutation rate in this example at $U_A = 5 \times 10^{-8}$. (D) Analytic predictions for the probability of observing a soft sweep in a sample of size two when the sweep starts at present (T_4), midway during the expansion ($T_3 = 50$ generations ago), at the beginning of the expansion ($T_2 = 113$ generations ago), and prior to the expansion ($T_1 = 500$ generations ago). Sweeps that start prior to the expansion are almost exclusively hard, whereas sweeps starting today will be soft in more than half of the cases, regardless of the strength of selection. Sweeps starting at the beginning or during the expansion show an interesting crossover behavior: smaller selection coefficients are more likely than larger selection coefficients to produce soft sweeps because weaker sweeps take longer to complete and thus experience more time at larger population sizes.